Tundra Swan State Space Model

Erik Osnas

18 May, 2017

I describe a Bayesian state-space model on which to base swan population assessment and as an alternative to a lagging three-year average of a population index. I also describe basic properties of Bayesian state-space models and how the model "smooths" the underlying state variables. While the three-year average is a straightforward method to smooth a variable index that constains sampling error, the rote use of three years is arbitrary. Why not a two- or five-year average? A two-year average is less smooth and a five-year average is more smooth. The appropriate amount of smoothing should depend on our understanding of the measurement error in the index, the true fluctuation in the population, and our knowledge of the underlying biological process. A Bayesian state-space model is a method to combine these to predict or estimate the underlying population state.

Swan Index Data

The swan index (Figure 1) is a summed composit of estimated swans from strata 8, 9, 10, and 11 from the Waterfowl Breeding Population and Habitat Survey (WBPHS) and the Yukon-Kuskokwim Delta Breeding Pair Survey (YKDBWS). The index is formed by summing swan estimates from the YKDBWS with strata 8, 10, 11 from the WBPHS and with estimates from the area of strata 9 of the WBPHS that does not overlap the area of the YKDBWS.

Note that some years have a high estimate followed by years with low estimates or vice versa (Figure 1). Many of the high estimates have corresponding high error estimates. This variation is a product of sampling (measurement) error and true underlying changes in the population from year-to-year in the sampled region (process error). Because the standard error estimates give the sampling error, we can use these in combination with a population model to estimate the process error and the underlying population state.

Fitting a Bayesian State Space Model

Here we fit a Bayesian state-space model to the swan data. A state-space model is a general term that usually refers to a model containing the true underlying (and unobserved) time-dependent state of the system and imperfect time-dependent observations. The states change with time so that the state at time t depends on the state at time $t-1$. Here, the state is defined as the unobserved total population of swans in the survey area, and observations depend just on the state. State-space models were first developed as part of missile guidance systems and the Apollo space program in the 1950s and 60s (the Kalman filter, see en.wikipedia.org/wiki/Kalman_filter), so in this sense a state-space model can be thought of as "rocket science." Ecological processes, however, are often more difficult to predict than rocket trajectories.

The model is Bayesian only because we implement it as such, but this also allows us to use prior information in the estimation of model parameters. You can think of prior information as our biological knowledge that is more general than any specific data set. You can also think of priors as a set of assumptions we make in order to use the methodology. When priors are based on highly specific and certain knowledge, they are called "informative" and can be very influential to the analysis. In contrast, "noninformative" priors could be used in cases of large uncertainty to simple bound the range of appropriate estimates. Priors are expressed as probability distribution with more "peaked" distributions being more informative than less "peaked" distributions.

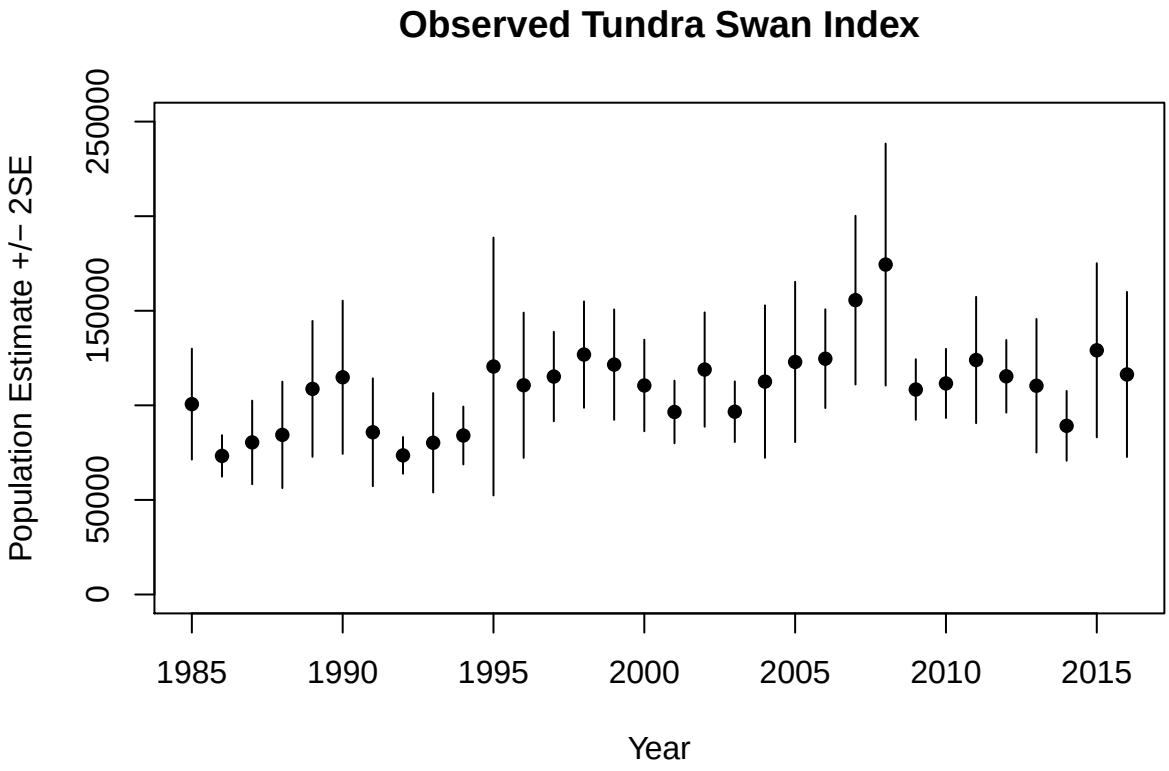

Figure 1: Swan population index from 1985 to 2016. Point estimates of the index are shown with the vertical bars representing 2 standard errors.

The model has four parts: a mathematical population projection model, an observation model, priors on all parameters, and observed data. The population projection model is

$N_{t+1} = N_t e^{r_t}$

with

$r_t \sim Normal(\bar{r}, \sigma_1)$.

N_t is the true population in the surveyed area (the state) and r_t are the year-specific change in the population on the log scale. The parameter $\bar{r}$ represents the long term average growth rate of the population on the log scale. The $\sim$ means "is distributed as" and signifies a random quantity. Thus, the underlying model is a stochastic exponential growth model. As σ_1 goes to zero, the population trajectory goes to the curve defined by $N_{t+1} = N_t e^{r_t}$; as σ_1 increases, the trajectory departs more from a fixed curve and looks more and more random as σ_1 increases further. These departures from the modeled curve are not "error" in the sense of random unrepeatable effects or measurement error, but represent true underlying departures from the mathematical model; therefore, this component of the variance is often called "process variance" or "process error" in the sense that these departure are "error" with respect to the mathematical model but still represent true population fluctuation from all causes not modeled. In order words, these deviations represent departures from a deterministic model.

The observation component of the model is

$Y_t \sim Normal(N_t, \sigma_{2,t})$

Here Y_t are the observed data (index values), and the standard deviation, $\sigma_{2,t}$, is indexed by time, meaning that it is year-specific. Because this is also calculated as part of the population index, $\sigma_{2,t}$ is input as data along with the index observations. The difference between N_t and Y_t represent errors in the sense of sampling noise–if the index survey was repeated with a different sample of transects and calculated anew (but all else being the same), the observed index value would be different. This error reflect our inability to observe the system perfectly. Having an estimate of the sampling variance (observation error) allows us to separate process variance from sample variance and to "smooth" the population estimate based on the data. The observation model defines the likelihood of the observing the data.

The final component of the model is a set of priors for the parameters. Priors represent our belief before we analyze the current data. One possible set of priors for the parameters is

$\bar{r} \sim Normal(0, 0.1)$,

$\sigma_1 \sim Uniform(0, 0.3)$, and

$log(N_1) \sim Normal(log(75000), 0.1)$.

The prior for $\bar{r}$ is normally distributed and centered on zero but could vary such that there is an approximately 95% probability that the true growth rate is $\pm$ 0.2 units from zero each year. At these extremes, the population is changing about $\pm$ 20% per year over more than 30 years. An average growth rate that is near $\pm$ 20% would mean the population is either quickly going extinct or will outgrow the available space on planet Earth. Clearly the true long term average growth rate is nearer to zero than these extremes, and this belief is reflected in the choice of a normal distribution with a mean of zero. The prior for parameter σ_1 is Uniform (equally probable) between 0 and 0.3 on the log population scale (standard deviation parameters can only be positive). This equates approximately to an annual coefficient of variation of about 30% at the maximum; meaning that if σ_1 is 0.3, then on average the annual deviation between the unobserved true population and the mathematical model will be about 30%. Any one deviation in a given year will be very different because it is a random quantity with a normal distribution on the log scale. On average across a large number of observations, however, deviations will be about 30% different from the mathematical model if σ_1 is set at the maximum. Finally, we need to specify a prior for the initial population size in the first year of observation. This basically constrains the model near biologically realistic populations sizes. Without this constraint, models of this type can sometimes be very difficult to fit. Here, the initial (log) population size is given a normal prior centered on 75000 swans and a coefficient of variation of about 10%. There is usually very little information to inform (estimate) this value in actual data sets.

Final estimates from a Bayesian analysis depend on both the priors and the data. When little or no information is contained in the data, estimates will to a greater degree reflect the prior. Conversely, when the data contributes much information, the final estimates (called "posterior estimates") can be very different from the priors. For time series-type data and when estimating variance components, as is the case here, 30 years of data can be thought of as mildly informative–not bad but also not great.

In the software language of BUGS (Lunn, et al. 2009) or JAGS (Plummer 2003) used through R (R Core Team 2015), the above model is represented as:

```
cat("model{
# Priors
  logN.est[1] ~ dnorm(log(75000), 1/(0.1)^2)  # Prior for initial population size log scale
  mean.r ~ dnorm(0, 1/(0.1)^2)                # Prior for mean growth rate
  sigma.proc ~ dunif(0, 0.3)                  # Prior for sd of state process log scale
  tau.proc <- pow(sigma.proc, -2)
# Likelihood
# State process
for (t in 1:(T-1) ) {
  r[t] ~ dnorm(mean.r, tau.proc)
  logN.est[t+1] <- logN.est[t] + r[t]
}
# Observation process
for (i in 1:T) {
  tau.obs[i] <- pow(sigma.obs[i], -2)
  y[i] ~ dnorm(exp(logN.est[i]), tau.obs[i])
}
}",
    file = "ssm.jags", fill = TRUE)
```

The above code writes a file to the directory, which is then used by BUGS or JAGS to define a simulation model that provides Bayesian posterior estimates. In R, we need to set up a list of data, a function that supplies initial values for each simulation, and then call and run the model. We run 3 independent simulations ("chains") that allow us to assess model convergence and be sure our estimates do not depend on the initial values supplied. In R using the package R2jags:

```
# structure data
jags.data <- list(T = length(dat$Year), y = dat$ALL,
    sigma.obs = dat$SE_ALL)
# Initial values
inits <- function() {
    list(logN.est = c(runif(1, log(74000), log(76000)),
        rep(NA, 31)), mean.r = runif(1, -1e-04, 1e-04),
        sigma.proc = runif(1, 0.01, 0.011), r = c(runif(31,
            -0.01, 0.01)))
}
# Parameters monitored
parameters <- c("logN.est", "mean.r", "sigma.proc")
# Call JAGS from R
out <- jags(jags.data, inits, parameters, model.file = "ssm.jags",
    n.chains = 3, n.thin = 1, n.iter = 11000, n.burnin = 10000,
    working.directory = getwd())
```

```
## Compiling model graph
##    Resolving undeclared variables
##    Allocating nodes
## Graph information:
```

```
##    Observed stochastic nodes: 32
##    Unobserved stochastic nodes: 34
##    Total graph size: 243
##
## Initializing model
```

A plot of some output and diagnostic information is examined to make sure the model worked (Figure 2). Inspecting the "R-hat" diagnostic, we see that it is near 1.0, which means that each chain is fluctuating around a similar distribution (the posterior).

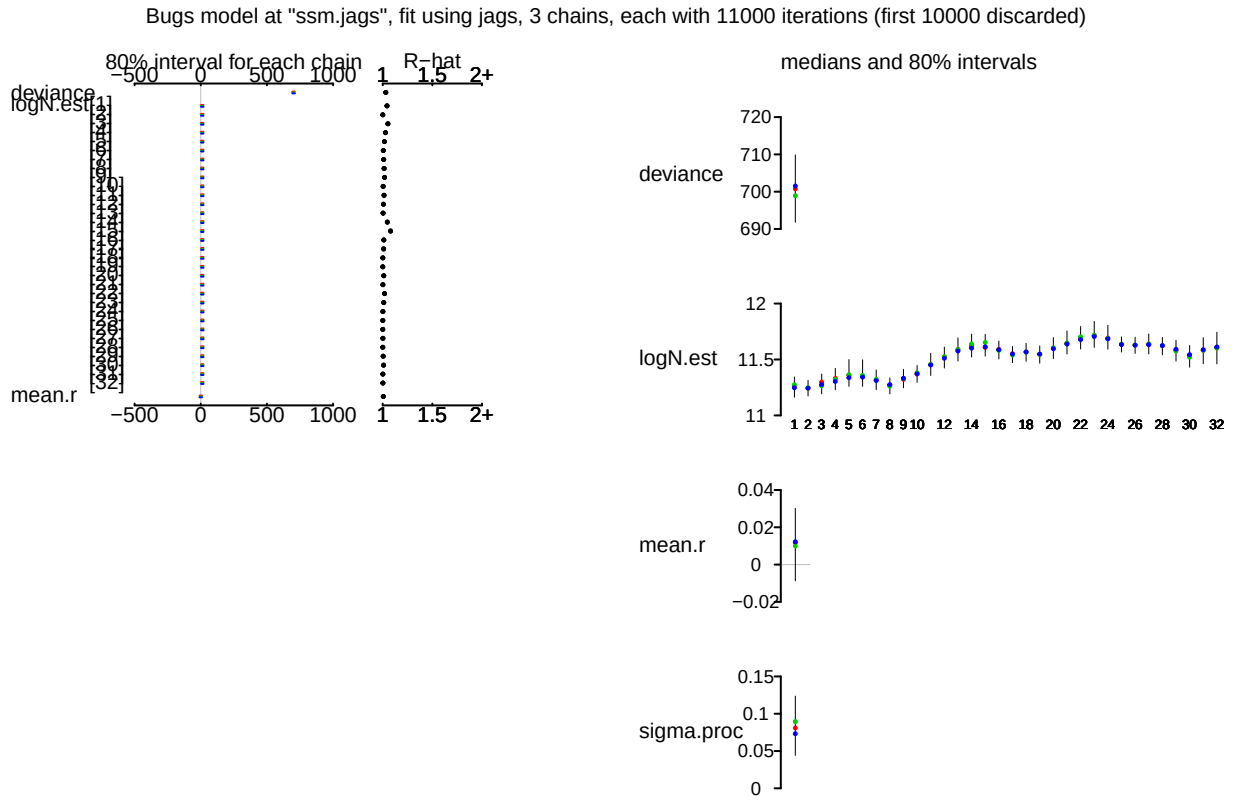

Figure 2: Model diagnoistic output to check convergence of the posterior simulations.

Now that we are confident that the model simulations settle on a posterior, we can plot the data and estimated parameters. The observed index values (data) with the posterior mean of the estimated population size (N_t) and the three-years average of the observed data are shown in Figure 3.

We can see that the posterior mean of the modeled population is less variable than the three-year average (Figure 3, Table 1). This will not be the case for all data sets. In this Bayesian model, the posterior mean population estimates are more smooth than the three-year average because the years with high point estimates for the index also have large standard errors (Figure 3). This causes the model to weight these data points less than others when estimating the modeled population. The degree of weight is determined by the observation variance that is input as data and on the estimated process variance. Because these parameters are estimated from the full time series of data and the priors, which themselves are based on our biological knowledge, the degree of smoothing is "based on all the data," explicit estimates of biological parameters (growth rate and it variance), and an explicit statement of our biological assumptions (the mathematical model and priors).

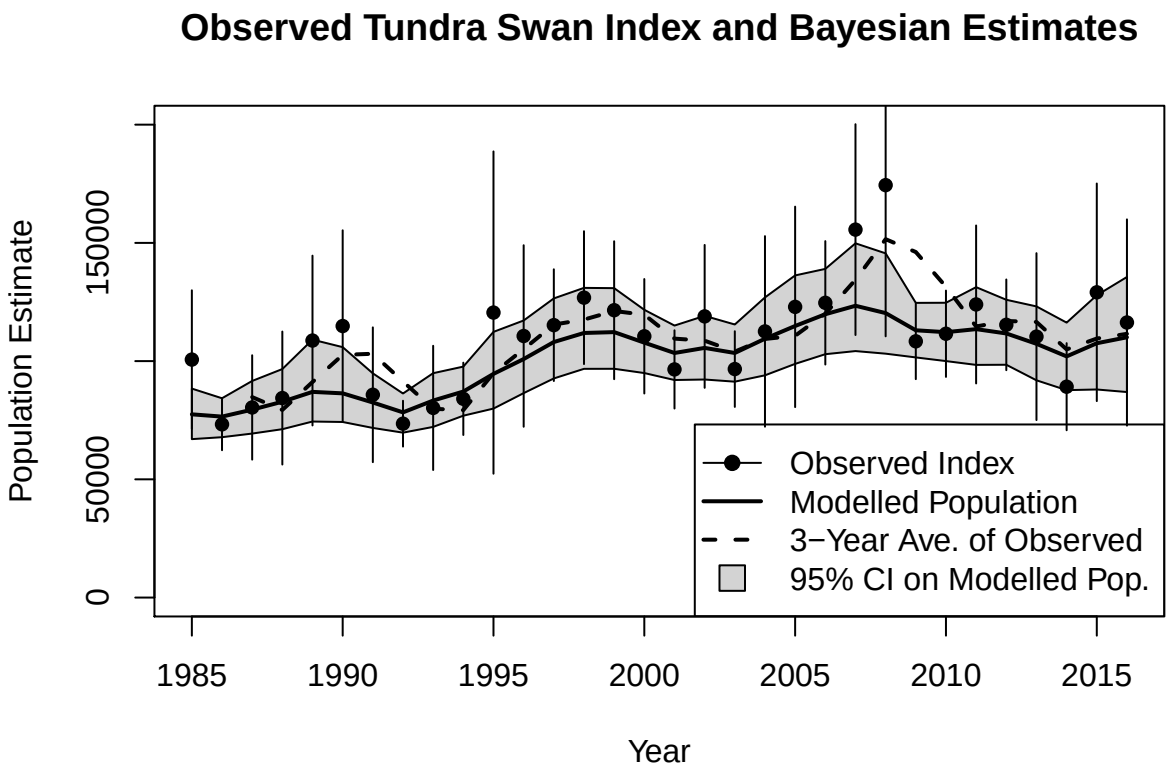

Figure 3: The observed swan index data plotted with the estimated posterior mean and 95% credible interval for the modelled population state and trailing three-year average of the observed index. The observed index is shown by the filled circles and vertical bars representing 2 standard errors of the index.

Table 1: The observed swan index (Observed Index), the standard deviation calculated from the index (SD Index), the posterior mean estimated population size from the Bayesian model described in the text (Posterior Mean Population), the standard deviation of the population size estimate (Posterior SD Population), and the trailing three year average of the observed index (Three Year Average).

Year	Observed Index	SD Index	Posterior Mean Population	Posterior SD Population	Three Year Average
1985	100645	14657	77490	5430	NA
1986	73264	5479	76528	4206	NA
1987	80404	11041	79478	5561	84771
1988	84379	14063	82827	6450	79349
1989	108692	17955	86984	8595	91158
1990	114834	20247	86365	8091	102635
1991	85764	14255	82498	5781	103097
1992	73539	4849	78320	4265	91379
1993	80211	13138	83319	5614	79838
1994	84049	7650	87128	5214	79266
1995	120529	34079	94603	7753	94930
1996	110623	19194	100866	7744	105067
1997	115213	11817	108063	8716	115455
1998	126846	14053	111894	8762	117561
1999	121525	14588	112258	8706	121195
2000	110506	12105	107807	6920	119626
2001	96491	8276	103445	5855	109507
2002	118919	15109	105584	6781	108639
2003	96636	8006	103436	6237	104015
2004	112556	20150	109462	8188	109370
2005	122939	21195	114856	9464	110710
2006	124651	13051	119828	9361	120049
2007	155615	22290	123401	11667	134402
2008	174428	31973	120316	10673	151565
2009	108369	8001	113036	5907	146137
2010	111575	9134	112240	6335	131457
2011	123962	16704	113527	8234	114635
2012	115345	9593	111697	6893	116961
2013	110358	17646	107316	7869	116555
2014	89177	9213	101994	7536	104960
2015	129102	23009	107561	9970	109546
2016	116328	21815	110068	12166	111536

Posterior estimates of two important biological parameters are shown as histograms for mean growth rate and process variance (Figure 4). We can see that the posterior for both mean growth rate and the process standard deviation is very different than the priors for each parameter (Figure 4). Recall that the prior for mean growth rate was Normal(0, 0.1) but that the posterior has a mean and standard deviation closer to 0.01. For the process variance, the prior was Uniform(0, 0.3); whereas, the posterior mean is 0.08 and the 97.5% quantile is 0.15.

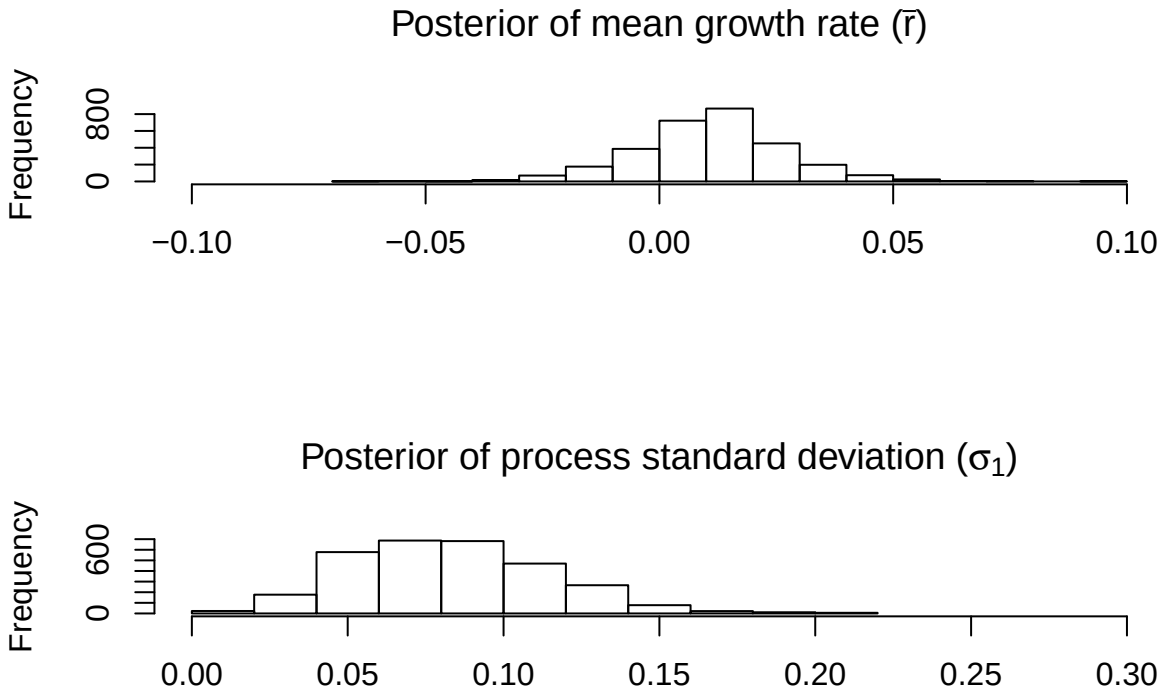

Figure 4: Posterior histograms for mean growth rate and process variance

Sensitivity Analysis on Priors

Here I vary the priors to see how large of an effect different priors might have. I only tried one different set that is much more informative for the mean growth rate and process variance and less informative for the initial population size. Before this model is adopted, additional thought should be given to priors.

```
cat("model{
# Priors
logN.est[1] ~ dnorm(log(75000), 1/(0.2)^2)  # Prior for initial population size log scale
mean.r ~ dnorm(0, 1/(0.05)^2)               # Prior for mean growth rate
sigma.proc ~ dunif(0, 0.1)                  # Prior for sd of state process log scale
tau.proc <- pow(sigma.proc, -2)
# Likelihood
# State process
for (t in 1:(T-1)){
  r[t] ~ dnorm(mean.r, tau.proc)
  logN.est[t+1] <- logN.est[t] + r[t]
}
# Observation process
for (i in 1:T) {
  tau.obs[i] <- pow(sigma.obs[i], -2)
  y[i] ~ dnorm(exp(logN.est[i]), tau.obs[i])
}
}",
    file = "ssm.2.jags", fill = TRUE)
```

```
# fit model to simulated data
jags.data <- list(T = length(dat$Year), y = dat$ALL,
    sigma.obs = dat$SE_ALL)
# Initial values
inits <- function() {
    list(logN.est = c(runif(1, log(74000), log(76000)),
        rep(NA, 31)), mean.r = runif(1, -1e-04, 1e-04),
        sigma.proc = runif(1, 0.01, 0.011), r = c(runif(31,
            -0.01, 0.01)))
}
# Parameters monitored
parameters <- c("logN.est", "mean.r", "sigma.proc")
# Call JAGS from R
out2 <- jags(jags.data, inits, parameters, model.file = "ssm.2.jags",
    n.chains = 3, n.thin = 1, n.iter = 11000, n.burnin = 10000,
    working.directory = getwd())
```

```
## Compiling model graph
##    Resolving undeclared variables
##    Allocating nodes
## Graph information:
##    Observed stochastic nodes: 32
##    Unobserved stochastic nodes: 34
##    Total graph size: 245
##
## Initializing model
```

Convergence is checked (Figure 5), and then the data and estimated parameters with new priors are plotted (Figure 6). As can be seen (Figure 6), there is very little effect of these different priors on the plotted results,

or the smoothing. The posterior histogram of mean growth rate is also similar to above, but process variance is truncated at 0.1 due to the prior (Figure 7). Despite this difference, there was very little practical effect on the smoothing.

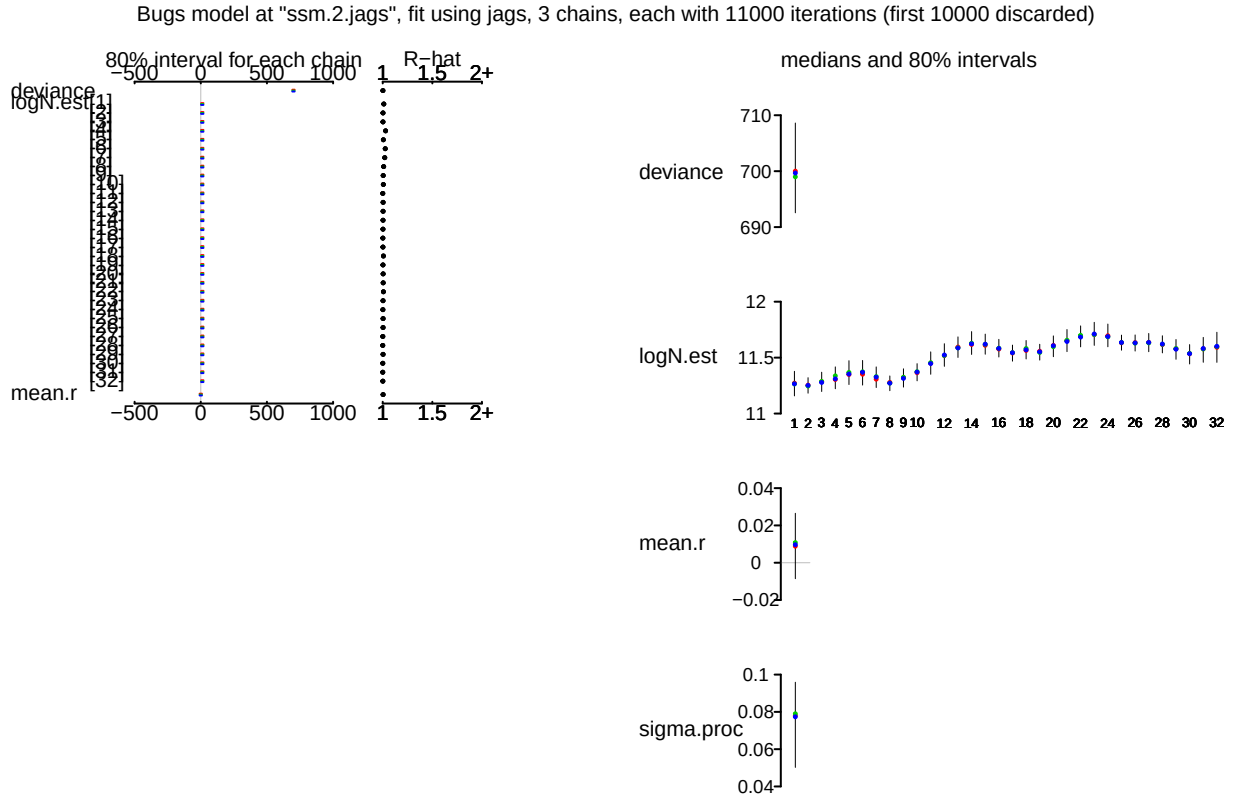

Figure 5: Model diagnoistic output to check convergence of the posterior simulations.

Discussion

Although Bayesian state-space models are more complex than calculating a lagging average, they provide an empirical and theoretical justification for the amount of index smoothing. Therefore, management decision rules can be based directly off the smoothed population index at the posterior mean or other posterior quantity. Despite the increased complexity, it is becoming routine to fit such model since the development of software programs in the 1990s (e.g., BUGS, JAGS, and now STAN). The current simulations took less than a minute to run per model and can easily be automated to produce results in a timely fashion for annual reports. More complex models could also be developed but at the expense of computation time and additional parameters, model assumptions, and priors.

One advantage of Bayesian state-space models not discussed above is the ability to estimate missing data. In the event that a survey is not completed, then the state-space model can predict what observation would have been made and the underlying state of the population, just as for other parameters. These estimates are based on the posterior distributions of all other parameters; therefore, the estimates of missing data are a product of the non-missing data, mathematical model, and the priors just as for other parameters. In addition, these predictions of missing data also contain the same degree of smoothing as the rest of the model and fully account for all uncertainties in parameter estimation. In the same way as missing data, predictions of the future population state can be made any number of years in the future.

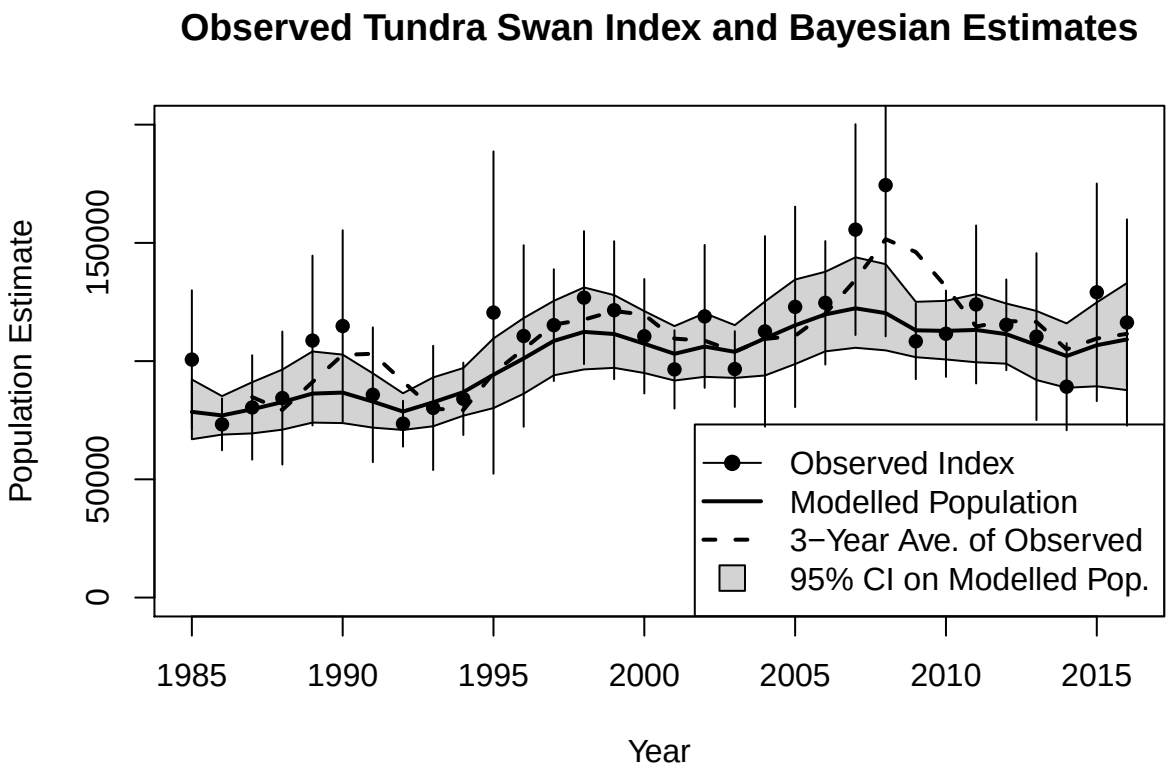

Figure 6: The observed swan index data plotted with the estimated posterior mean and 95% credible interval for the modelled population state and trailing three-year average of the observed index. The observed index is shown by the filled circles and vertical bars representing 2 standard errors of the index.

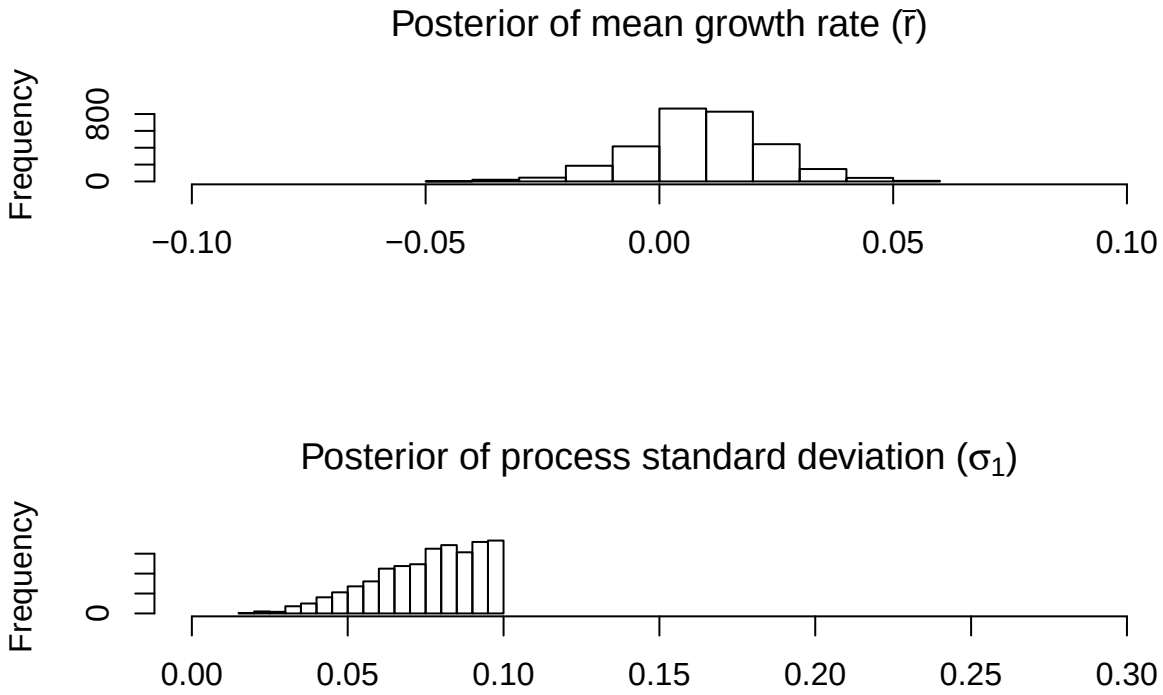

Figure 7: Posterior histograms for mean growth rate and process variance

With respect to missing data, another model that might be considered is fitting a state-space prediction separately to each component of the index. That is, a state-space model as above could be fit to each strata component and then this space-space parameter is summed across all strata. Each strata data stream would have a separate process variance, observation variance, and mean growth rate. This would introduce additional complexity but would allow missing data to be modeled at the strata level rather than the summed total of all strata. In the case that a survey or strata could not be complete, uncertainty in predicting missing data would be restricted to that strata while other data could still be used. Additionally, strata-specific estimate of population size and growth rate would be available. Whether or not these advantages outweigh the additional model complexity should be considered. If missing data are unlikely or strata species estimate are not desired, then a simpler model would be warranted.

References

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Lunn, D.; Spiegelhalter, D.; Thomas, A.; Best, N. (2009). "The BUGS project: Evolution, critique and future directions". Statistics in Medicine. 28 (25): 3049-3067.

Plummer, M. (2003). JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling, Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003), March 20-22, Vienna, Austria.